PACLIC 28

Discourse for Machine Translation

Bonnie Webber
School of Informatics, University of Edinburgh

Abstract

Statistical Machine Translation is a modern success: Given a source language sentence, SMT finds the most probable target language sentence, based on (1) properties of the source; (2) probabilistic source--target mappings at the level of words, phrases and/or sub-structures; and (3) properties of the target language.

SMT translates individual sentences because the search space even for a single sentence can be vast. But sentences are parts of texts, and texts have properties beyond those of their individual sentences, including:

- document-wide properties, such as style, register, reading level and genre, that are visible in the frequency and distribution of words, word senses, referential forms and syntactic structures;
- patterns of topical or functional sub-structures that mean that frequencies and distributions of words, word senses, referential forms and syntactic structures will vary across a text;
- relations between clauses or between referring expressions that can be signaled explicitly or implicitly, that reflect a text's coherence;
- frequent appeal to reduced expressions that rely on context to
- efficiently convey their message.

Recognizing and deploying these properties promises to improve both fluency and accuracy in SMT -- i.e., whether the sequence of sentences in the target text conveys the same information as those in its source, in as readable a manner. This presentation describes how researchers are attempting to do this, without bringing translation to a halt.